CLAIMS

What is claimed is:

1. A method for automatically identifying relationships between text documents and structured variables pertaining to said text documents, comprising:

generating a dictionary of keywords in said text documents;

forming categories of said text documents using said dictionary and an automated algorithm;

counting occurrences of said structured variables, said categories and said structured variable/category combinations in said text documents; and

calculating probabilities of occurrences of said structured variable/category combinations.

2. The method according to claim 1, wherein said algorithm comprises a keyword occurrence algorithm and wherein each of said categories comprises a category of text documents in which a particular keyword occurs.

3. The method according to claim 1, wherein said algorithm comprises a clustering algorithm and wherein each of said categories comprises a category of said text documents containing a particular cluster.

4. The method according to claim 3, wherein said clustering algorithm comprises a k means

algorithm.

5. The method according to claim 3, wherein said forming categories comprises inputting a predetermined number of categories.

6. The method according to claim 2, wherein said forming categories comprises:

generating a sparse matrix array containing a count of each of said keywords in each of said text documents.

7. The method according to claim 1, wherein said keywords comprise words or phrases which occur a predetermined number of times in said text documents.

8. The method of claim 1, wherein said calculating probabilities comprises using a Chi squared function.

9. The method of claim 1, wherein said generating a dictionary of keywords comprises:

first parsing text in said text document to identify and count occurrences of words;

storing a predetermined number of frequently occurring words;

second parsing text in said text documents to identify and count occurrences of phrases; and

storing a predetermined number of frequently occurring phrases.

10. The method according to claim 9, wherein said frequently occurring words and phrases are stored in a hash table.

11. The method according to claim 6, wherein said generating a sparse matrix array comprises:

third parsing text in said text documents to count a number of times that each of said keywords occurs in each of said text documents.

12. The method according to claim 1, wherein said relationships comprise structured variable/category combinations having a lowest probability of occurrence.

13. The method according to claim 1, wherein said method comprises a computer implemented method.

14. The method according to claim 1, wherein said method calculates a probability that a given co-occurrence of a structured variable and a category would have occurred as a purely random event.

15. The method according to claim 1, wherein said structured variables comprise predetermined time intervals.

16. The method according to claim 15, wherein said predetermined time intervals comprise one of days, weeks, months and years.

17. A system for automatically identifying relationships between text documents and structured variables pertaining to said text documents, comprising:

an input device for inputting text documents;

a processor for forming categories of said text documents and counting occurrences of said structured variables, categories and structured variable/category combinations and calculating probabilities of occurrence of said structured variable/category combinations; and

a display, for displaying said probabilities.

18. The system according to claim 17, further comprising:

a memory for storing occurrences of said structured variables, categories and structured variable/category combinations and probabilities of occurrences of said structured variable/category combinations.

19. The system according to claim 17, wherein said structured variables comprise predetermined time intervals.

20. The system according to claim 19, wherein said predetermined time intervals comprise one of days, weeks, months and years.

21. The system according to claim 17, wherein said system calculates a probability that a given co-occurrence of a structured variable and a category would have occurred as a purely random event.

22. The system according to claim 17, wherein said relationships comprise statistically significant relationships.

23. A programmable storage medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method for automatically identifying relationships between text documents and structured variables pertaining to said text documents, said method comprising:

generating a dictionary of keywords in said text documents;

forming categories of said text documents using said dictionary and an automated algorithm;

counting occurrences of said structured variables, said categories and said structured variable/category combinations in said text documents; and

calculating probabilities of occurrences of said structured variable/category combinations.